BIOINFORMATICS
Introduction

Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mbb452a
Bioinformatics

Biological Data + Computer Calculations
What is Bioinformatics?

• *(Molecular)* **Bio-informatics**

• One idea for a definition? Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **informatics’ techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale**.

• Bioinformatics is “MIS” for Molecular Biology Information. It is a practical discipline with many **applications**.
What is the **Information**?

Molecular Biology as an Information Science

- **Central Dogma of Molecular Biology**
  - DNA
    - RNA
    - Protein
    - Phenotype
    - DNA

- **Molecules**
  - Sequence, Structure, Function

- **Processes**
  - Mechanism, Specificity, Regulation

- **Central Paradigm for Bioinformatics**
  - Genomic Sequence Information
    - mRNA (level)
    - Protein Sequence
    - Protein Structure
    - Protein Function
    - Phenotype

- **Large Amounts of Information**
  - Standardized
  - Statistical

- **Genetic material**
  - Information transfer (mRNA)
  - Protein synthesis (tRNA/mRNA)
  - Some catalytic activity

- Most cellular functions are performed or facilitated by proteins.
  - Primary biocatalyst
  - Cofactor transport/storage
  - Mechanical motion/support
  - Immune protection
  - Control of growth/differentiation

(idea from D Brutlag, Stanford, graphics from S Strobel)
Molecular Biology Information - DNA

- Raw DNA Sequence
  - Coding or Not?
  - Parse into genes?
  - 4 bases: AGCT
  - ~1 K in a gene, ~2 M in genome

atggcaat...
Molecular Biology Information: Protein Sequence

• 20 letter alphabet
  ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ

• Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain

• ~200 K known protein sequences

| d1dhfa_ | LNCIVAVSQNMIGKNGDLPMPLRNLFYFQORMTSSVEGKQ-NLVIMGKKTWSFI |
| d8dfr__ | LNSIVAVCNMIGKDNGLPNPLRNKYYFQORMTSTSHVEGKQ-NAVIMGKKTWSFI |
| d4dfra__ | ISLIALAVDRVSQMNIMAPNLADLFKRNTL----------NPVIMGREHTWSI |
| d3dfr__  | TAFLWASDRDGGLGHDGHLPHLDDLLHYFRAQTV---------GKIMVGVRTYEF |

| d1dhfa_ | LNCIVAVSQNMIGKNGDLPMPLRNLFYFQORMTSSVEGKQ-NLVIMGKKTWSFI |
| d8dfr__ | LNSIVAVCNMIGKDNGLPNPLRNKYYFQORMTSTSHVEGKQ-NAVIMGKKTWSFI |
| d4dfra__ | ISLIALAVDRVSQMNIPNLADLFKRNTL----------KPVIMGREHTWSI |
| d3dfr__  | TAFLWASDRDGGLGHDGHLPHLDDLLHYFRAQTV---------KIMVGVRTYEF |

| d1dhfa_ | VPEKNRPLKRGIVLSSRKELPQGAFFLSRSLDLLKLETQELANKVDMVWIVGSSVYEKAMNH |
| d8dfr__ | VPEKNRPLKRGIVLSSRKELPQGAFFLSRSLDLLKLETQELANKVDMVWIVGSSVYEKAMNH |
| d4dfra__ | ---G-RPLGRKNIILSSQPGLDDRVTWKSVDDEAIACGDPV-----EMVIGGVRVYEQFPLKA |
| d3dfr__  | ---PKRPFLERTNVLTHQEDYQAQGA-VVHDVAAVVFAYAKHLDQ-----ELVIAAGQIFTAFKDDV |

| d1dhfa_ | -PEKNRPLKRGIVLSSRKELPQGAFFLSRSLDLLKLETQELANKVDMVWIVGSSVYEKAMNH |
| d8dfr__ | -PEKNRPLKRGIVLSSRKELPQGAFFLSRSLDLLKLETQELANKVDMVWIVGSSVYEKAMNH |
| d4dfra__ | -G-RRPLGRKNIILSSQPGLDDRVTWKSVDDEAIACGDPV-----EMVIGGVRVYEQFPLKA |
| d3dfr__  | -P---PKRPFLERTNVLTHQEDYQAQGA-VVHDVAAVVFAYAKHLDQ-----ELVIAAGQIFTAFKDDV |
Molecular Biology Information: Macromolecular Structure

- DNA/RNA/Protein
  - Almost all protein

(Adapted From D. Soll Web Page, Right Hand Top Protein from M. Levitt web page)
Molecular Biology Information: Protein Structure Details

• Statistics on Number of XYZ triplets
  ◊ 200 residues/domain -> 200 CA atoms, separated by 3.8 A
  ◊ Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic A
  ◊ => ~1500 xyz triplets (=8x200) per protein domain
  ◊ 10 K known domain, ~300 folds

ATOM     1  C   ACE     0       9.401  30.166  60.595  1.00 49.88      1GKY  67
ATOM     2  O   ACE     0      10.432  30.832  60.722  1.00 50.35      1GKY  68
ATOM     3  CH3 ACE      0       8.876  29.767  59.226  1.00 50.04      1GKY  69
ATOM     4  N   SER      1       8.753  29.755  61.685  1.00 49.13      1GKY  70
ATOM     5  CA  SER      1       9.242  30.200  62.974  1.00 46.62      1GKY  71
ATOM     6  C   SER      1      10.453  29.500  63.579  1.00 41.99      1GKY  72
ATOM     7  O   SER      1      10.593  29.607  64.814  1.00 43.24      1GKY  73
ATOM     8  CB  SER      1       8.052  30.189  63.974  1.00 53.00      1GKY  74
ATOM     9  OG  SER      1       7.294  31.409  63.930  1.00 57.79      1GKY  75
ATOM    10  N   ARG      2      11.360  28.819  62.827  1.00 36.48      1GKY  76
ATOM    11  CA  ARG      2      12.548  28.316  63.532  1.00 25.54      1GKY  77
ATOM    12  C   ARG      2      13.502  29.501  63.500  1.00 25.54      1GKY  78

ATOM   1444  CB  LYS   186     13.836  22.263  57.567  1.00 55.06      1GKY1510
ATOM   1445  CG  LYS   186     12.422  22.452  58.180  1.00 53.45      1GKY1511
ATOM   1446  CD  LYS   186     11.531  21.198  58.185  1.00 49.88      1GKY1512
ATOM   1447  CE  LYS   186     11.452  20.402  56.860  1.00 48.15      1GKY1513
ATOM   1448  NZ  LYS   186     10.735  21.104  55.811  1.00 48.41      1GKY1514
ATOM   1449  OXT LYS   186     16.887  23.841  56.647  1.00 62.94      1GKY1515
TER   1450  LYS   186

ATOM   1444  CB  LYS   186    3.8 A

3.8 A
Molecular Biology
Information: Whole Genomes

• The Revolution Driving Everything


(Picture adapted from TIGR website, http://www.tigr.org)

• Integrative Data

1995, HI (bacteria): 1.6 Mb & 1600 genes done
1997, yeast: 13 Mb & ~6000 genes for yeast
1998, worm: ~100Mb with 19 K genes
1999: >30 completed genomes!
2003, human: 3 Gb & 100 K genes...

Genome sequence now accumulate so quickly that, in less than a week, a single laboratory can produce more bits of data than Shakespeare managed in a lifetime, although the latter make better reading.

Genomes highlight the Finiteness of the “Parts” in Biology

1995
Bacteria, 1.6 Mb, ~1600 genes
[Science 269: 496]

1997
Eukaryote, 13 Mb, ~6K genes
[Nature 387: 1]

1998
Animal, ~100 Mb, ~20K genes
[Science 282: 1945]

2000?
Human, ~3 Gb, ~100K genes [??]

real thing, Apr ’00

‘98 spoof
Gene Expression Datasets: the Transcriptosome

Young/Lander, Chips, Abs. Exp.

Also: SAGE, Samson and Church, Chips;
Protein Expression, Aebersold, Protein Expression

Transposons, Snyder, Transposons, Protein Exp.

Brown, microarray, Rel. Exp. over Timecourse

Gene Expression

Dissecting the Regulatory Circuity of a Eukaryotic Genome

Brown/Lander, Chips, Abs. Exp.
Array Data

**Yeast Expression Data in Academia:**
levels for all 6000 genes!

Can only sequence genome once but can do an infinite variety of these array experiments

at 10 time points,
6000 x 10 = 60K floats

telling signal from background

(courtesy of J Hager)
Other Whole-Genome Experiments

Systematic Knockouts


2 hybrids, linkage maps


For yeast: 6000 x 6000 / 2 ~ 18M interactions
Molecular Biology Information: Other Integrative Data

- Information to understand genomes
  - Metabolic Pathways (glycolysis), traditional biochemistry
  - Regulatory Networks
  - Whole Organisms Phylogeny, traditional zoology
  - Environments, Habitats, ecology
  - The Literature (MEDLINE)

- The Future....

(Pathway drawing from P Karp’s EcoCyc, Phylogeny from S J Gould, Dinosaur in a Haystack)
What is Bioinformatics?

• (Molecular) **Bio-informatics**

• One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **informatics** techniques (derived from disciplines such as applied math, CS, and statistics) to understand and organize the **information associated** with these molecules, **on a large-scale**.

• Bioinformatics is “MIS” for Molecular Biology Information. It is a practical discipline with many **applications**.
### GenBank Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Base Pairs</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>680338</td>
<td>606</td>
</tr>
<tr>
<td>1983</td>
<td>2274029</td>
<td>2427</td>
</tr>
<tr>
<td>1984</td>
<td>3368765</td>
<td>4175</td>
</tr>
<tr>
<td>1985</td>
<td>5204420</td>
<td>5700</td>
</tr>
<tr>
<td>1986</td>
<td>9615371</td>
<td>9978</td>
</tr>
<tr>
<td>1987</td>
<td>15514776</td>
<td>14584</td>
</tr>
<tr>
<td>1988</td>
<td>23800000</td>
<td>20579</td>
</tr>
<tr>
<td>1989</td>
<td>34762585</td>
<td>28791</td>
</tr>
<tr>
<td>1990</td>
<td>49179285</td>
<td>39533</td>
</tr>
<tr>
<td>1991</td>
<td>71947426</td>
<td>55627</td>
</tr>
<tr>
<td>1992</td>
<td>101008486</td>
<td>78608</td>
</tr>
<tr>
<td>1993</td>
<td>157152442</td>
<td>143492</td>
</tr>
<tr>
<td>1994</td>
<td>217102462</td>
<td>215273</td>
</tr>
<tr>
<td>1995</td>
<td>384939485</td>
<td>555694</td>
</tr>
<tr>
<td>1996</td>
<td>651972984</td>
<td>1021211</td>
</tr>
<tr>
<td>1997</td>
<td>1160300687</td>
<td>1765847</td>
</tr>
<tr>
<td>1998</td>
<td>2008761784</td>
<td>2837897</td>
</tr>
<tr>
<td>1999</td>
<td>3841163011</td>
<td>4864570</td>
</tr>
<tr>
<td>2000</td>
<td>8604221980</td>
<td>7077491</td>
</tr>
</tbody>
</table>
Large-scale Information: Explonential Growth of Data Matched by Development of Computer Technology

- CPU vs Disk & Net
  - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Driving Force in Bioinformatics

(Internet picture adapted from D Brutlag, Stanford)
Growth in number of residues in Genbank, a central database for sequence data, compared to the request for people with competence in bioinformatics. The request for scientists is estimated from the number of relevant positions advertised in the first number of Nature in March and September of each year.

(courtesy of Finn Drablos)
"Don’t just sit there! If you’ve processed all the data there is, go out and find more data!"

What is Bioinformatics?

• (Molecular) **Bio-informatics**

• One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying “informatics” **techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale**.

• Bioinformatics is “MIS” for Molecular Biology Information. It is a practical discipline with many **applications**.
Organizing Molecular Biology Information: Redundancy and Multiplicity

- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genes are grouped into Pathways
- Genomic Sequence Redundancy due to the Genetic Code
- How do we find the similarities?.....
Molecular Parts = Conserved Domains, Folds, &c
A Parts List Approach to Bike Maintenance
A Parts List Approach to Bike Maintenance

What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts - - types of parts (nuts & washers)?

How many roles can these play? How flexible and adaptable are they mechanically?

Where are the parts located?
Vast Growth in (Structural) Data... but number of Fundamentally New (Fold) Parts Not Increasing that Fast.
World of Structures is even more Finite, providing a valuable simplification

(T. pallidum)

Same logic for pathways, functions, sequence families, blocks, motifs....

Global Surveys of a Finite Set of Parts from Many Perspectives

Functions picture from www.fruitfly.org/~suzi (Ashburner); Pathways picture from, ecocyc.pangeaSystems.com/ecocyc (Karp, Riley). Related resources: COGS, ProDom, Pfam, Blocks, Domo, WIT, CATH, Scop....
What is Bioinformatics?

- (Molecular) Bio-informatics
- One idea for a definition? Bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying “informatics” techniques (derived from disciplines such as applied math, CS, and statistics) to understand and organize the information associated with these molecules, on a large-scale.
- Bioinformatics is “MIS” for Molecular Biology Information. It is a practical discipline with many applications.
General Types of “Informatics” techniques in Bioinformatics

- Databases
  - Building, Querying
  - Object DB
- Text String Comparison
  - Text Search
  - 1D Alignment
  - Significance Statistics
  - Alta Vista, grep
- Finding Patterns
  - AI / Machine Learning
  - Clustering
  - Datamining
- Geometry
  - Robotics
  - Graphics (Surfaces, Volumes)
  - Comparison and 3D Matching (Vision, recognition)
- Physical Simulation
  - Newtonian Mechanics
  - Electrostatics
  - Numerical Algorithms
  - Simulation
New Paradigm for Scientific Computing

• Because of increase in data and improvement in computers, new calculations become possible

• But Bioinformatics has a new style of calculation...

  ◊ Two Paradigms

• Physics

  ◊ Prediction based on physical principles
  ◊ Exact Determination of Rocket Trajectory
  ◊ Supercomputer, CPU

• Biology

  ◊ Classifying information and discovering unexpected relationships
  ◊ globin ~ colicin ~ plastocyanin ~ repressor
  ◊ networks, “federated” database
Bioinformatics Topics --
Genome Sequence

• Finding Genes in Genomic DNA
  ◊ introns
  ◊ exons
  ◊ promotors
• Characterizing Repeats in Genomic DNA
  ◊ Statistics
  ◊ Patterns
• Duplications in the Genome
• Sequence Alignment
  ◊ non-exact string matching, gaps
  ◊ How to align two strings optimally via Dynamic Programming
  ◊ Local vs Global Alignment
  ◊ Suboptimal Alignment
  ◊ Hashing to increase speed (BLAST, FASTA)
  ◊ Amino acid substitution scoring matrices

• Multiple Alignment and Consensus Patterns
  ◊ How to align more than one sequence and then fuse the result in a consensus representation
  ◊ Transitive Comparisons
  ◊ HMMs, Profiles
  ◊ Motifs

Bioinformatics
Topics --
Protein Sequence

• Scoring schemes and Matching statistics
  ◊ How to tell if a given alignment or match is statistically significant
  ◊ A P-value (or an e-value)?
  ◊ Score Distributions (extreme val. dist.)
  ◊ Low Complexity Sequences
Bioinformatics

Topics --
Sequence / Structure

• Secondary Structure “Prediction”
  ◊ via Propensities
  ◊ Neural Networks, Genetic Alg.
  ◊ Simple Statistics
  ◊ TM-helix finding
  ◊ Assessing Secondary Structure Prediction

• Tertiary Structure Prediction
  ◊ Fold Recognition
  ◊ Threading
  ◊ Ab initio

• Function Prediction
  ◊ Active site identification

• Relation of Sequence Similarity to Structural Similarity
Topics -- Structures

• Basic Protein Geometry and Least-Squares Fitting
  ◊ Distances, Angles, Axes, Rotations
  • Calculating a helix axis in 3D via fitting a line
  ◊ LSQ fit of 2 structures
  ◊ Molecular Graphics

• Calculation of Volume and Surface
  ◊ How to represent a plane
  ◊ How to represent a solid
  ◊ How to calculate an area
  ◊ Docking and Drug Design as Surface Matching
  ◊ Packing Measurement

• Structural Alignment
  ◊ Aligning sequences on the basis of 3D structure.
  ◊ DP does not converge, unlike sequences, what to do?
  ◊ Other Approaches: Distance Matrices, Hashing
  ◊ Fold Library
• Relational Database Concepts
  ◊ Keys, Foreign Keys
  ◊ SQL, OODBMS, views, forms, transactions, reports, indexes
  ◊ Joining Tables, Normalization
    • Natural Join as "where" selection on cross product
    • Array Referencing (perl/dbm)
  ◊ Forms and Reports
  ◊ Cross-tabulation
• Protein Units?
  ◊ What are the units of biological information?
    • sequence, structure
    • motifs, modules, domains
  ◊ How classified: folds, motions, pathways, functions?

Topics -- Databases

• Clustering and Trees
  ◊ Basic clustering
    • UPGMA
    • single-linkage
    • multiple linkage
  ◊ Other Methods
    • Parsimony, Maximum likelihood
  ◊ Evolutionary implications

• The Bias Problem
  ◊ sequence weighting
  ◊ sampling
Topics -- Genomics

- Expression Analysis
  ◦ Time Courses clustering
  ◦ Measuring differences
  ◦ Identifying Regulatory Regions
- Large scale cross referencing of information
- Function Classification and Orthologs
- The Genomic vs. Single-molecule Perspective

- Genome Comparisons
  ◦ Ortholog Families, pathways
  ◦ Large-scale censuses
  ◦ Frequent Words Analysis
  ◦ Genome Annotation
  ◦ Trees from Genomes
  ◦ Identification of interacting proteins

- Structural Genomics
  ◦ Folds in Genomes, shared & common folds
  ◦ Bulk Structure Prediction

- Genome Trees
Topics -- Simulation

• Molecular Simulation
  ◊ Geometry → Energy → Forces
  ◊ Basic interactions, potential energy functions
  ◊ Electrostatics
  ◊ VDW Forces
  ◊ Bonds as Springs
  ◊ How structure changes over time?
    • How to measure the change in a vector (gradient)
  ◊ Molecular Dynamics & MC
  ◊ Energy Minimization

• Parameter Sets
• Number Density
• Poisson-Boltzman Equation
• Lattice Models and Simplification
### Bioinformatics Schematic

<table>
<thead>
<tr>
<th>Depth:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rational</td>
</tr>
<tr>
<td>Drug</td>
</tr>
<tr>
<td>Design, physics</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Breadth:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homologues, Large-scale Surveys, informatics</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pairwise comparison, sequence &amp; structure alignment</th>
<th>multiple alignment, patterns, templates, trees</th>
<th>databases, scoring schemes, censuses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome Sequence</td>
<td>g₁ ↔ g₂</td>
<td>g₁ ↔ g₂, g₃, g₄... g₅₀ ↔ g₅₀... g₁₀₀₀...</td>
</tr>
<tr>
<td>Protein Sequence</td>
<td>p₁ ↔ p₂</td>
<td>p₁ ↔ p₂, p₃, p₄... p₅₀ ↔ p₅₀... p₁₀₀₀...</td>
</tr>
<tr>
<td>structure prediction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein Structure</td>
<td>s₁ ↔ s₂</td>
<td>s₁ ↔ s₂, s₃, s₄... s₅₀ ↔ s₅₀... s₁₀₀₀...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>geometry calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Surface</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>molecular simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Force Field</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>structure docking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ligand Complex</td>
</tr>
</tbody>
</table>
# Background

<table>
<thead>
<tr>
<th>Need to Know Today</th>
<th>Math</th>
<th>Biology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation of Standard Deviation, a Bell-shaped Distribution (of test scores), a 3D vector</td>
<td></td>
<td>DNA, RNA, alpha-helix, the cell nucleus, ATP</td>
</tr>
<tr>
<td>What You’ll Learn</td>
<td>Force is the Derivative (grad) of Energy, Rotation Matrices (3D), a P-value of .01 and an Extreme Value Distribution</td>
<td>Proteins are tightly packed, sequence homology twilight zone, protein families</td>
</tr>
<tr>
<td>Not really necessary….</td>
<td>Poisson-Boltzman Equation, Design a Hashing Function, Write a Recursive Descent Parser</td>
<td>What GroEL does, a worm is a metazoa, E. coli is gram negative, what chemokines are</td>
</tr>
</tbody>
</table>
Are They or Aren’t They Bioinformatics? (#1)

• Digital Libraries
  ◊ Automated Bibliographic Search and Textual Comparison
  ◊ Knowledge bases for biological literature

• Motif Discovery Using Gibb's Sampling

• Methods for Structure Determination
  ◊ Computational Crystallography
    • Refinement
  ◊ NMR Structure Determination
    • Distance Geometry

• Metabolic Pathway Simulation

• The DNA Computer
Are They or Aren’t They Bioinformatics? (#1, Answers)

- (YES?) Digital Libraries
  ◊ Automated Bibliographic Search and Textual Comparison
  ◊ Knowledge bases for biological literature
- (YES) Motif Discovery Using Gibb's Sampling
- (NO?) Methods for Structure Determination
  ◊ Computational Crystallography
    • Refinement
  ◊ NMR Structure Determination
    • (YES) Distance Geometry
- (YES) Metabolic Pathway Simulation
- (NO) The DNA Computer
Are They or Aren’t They Bioinformatics? (#2)

- Gene identification by sequence inspection
  ◊ Prediction of splice sites
- DNA methods in forensics
- Modeling of Populations of Organisms
  ◊ Ecological Modeling
- Genomic Sequencing Methods
  ◊ Assembling Contigs
  ◊ Physical and genetic mapping
- Linkage Analysis
  ◊ Linking specific genes to various traits
Are They or Aren’t They Bioinformatics? (#2, Answers)

- **(YES)** Gene identification by sequence inspection
  ◊ Prediction of splice sites
- **(YES)** DNA methods in forensics
- **(NO)** Modeling of Populations of Organisms
  ◊ Ecological Modeling
- **(NO?)** Genomic Sequencing Methods
  ◊ Assembling Contigs
  ◊ Physical and genetic mapping
- **(YES)** Linkage Analysis
  ◊ Linking specific genes to various traits
Are They or Aren’t They Bioinformatics? (#3)

• RNA structure prediction
  Identification in sequences
• Radiological Image Processing
  ◊ Computational Representations for Human Anatomy (visible human)
• Artificial Life Simulations
  ◊ Artificial Immunology / Computer Security
  ◊ Genetic Algorithms in molecular biology
• Homology modeling
• Determination of Phylogenies Based on Non-molecular Organism Characteristics
• Computerized Diagnosis based on Genetic Analysis (Pedigrees)
Are They or Aren’t They Bioinformatics? (#3, Answers)

• (YES) RNA structure prediction
  Identification in sequences

• (NO) Radiological Image Processing
  ◊ Computational Representations for Human Anatomy (visible human)

• (NO) Artificial Life Simulations
  ◊ Artificial Immunology / Computer Security
  ◊ (NO?) Genetic Algorithms in molecular biology

• (YES) Homology modeling

• (NO) Determination of Phylogenies Based on Non-molecular Organism Characteristics

• (NO) Computerized Diagnosis based on Genetic Analysis (Pedigrees)
What is Bioinformatics?

• (Molecular) **Bio-informatics**

• One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **“informatics” techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale**.

• Bioinformatics is “MIS” for Molecular Biology Information. It is a practical discipline with many **applications**.
**Major Application I:** Designing Drugs

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).
Major Application II: Finding Homologs

1. ISOLATE HUMAN DNA SEQUENCE

2. TRANSLATE DNA SEQUENCE INTO AMINO ACID SEQUENCES

3. FIND SIMILAR SEQUENCES IN DATABASES OF MODEL ORGANISM PROTEINS (green areas reflect great differences; orange, smaller variations)

4. MODEL HUMAN PROTEIN STRUCTURE BASED ON KNOWN STRUCTURE OF A SIMILAR PROTEIN FROM A MODEL ORGANISM (red area is encoded by the sequence shown)

5. FIND DRUG THAT BINDS TO MODELED PROTEIN
Major Application II: Finding Homologues

- Find Similar Ones in Different Organisms
- Human vs. Mouse vs. Yeast
  ◊ Easier to do Expts. on latter!

(Section from NCBI Disease Genes Database Reproduced Below.)

<table>
<thead>
<tr>
<th>Human Disease</th>
<th>MIM #</th>
<th>Human Gene</th>
<th>GenBank Acc# for P-value</th>
<th>Human GenBank</th>
<th>BLASTX Gene</th>
<th>Yeast GenBank</th>
<th>Yeast Gene Acc# for Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hereditary Non-polyposis Colon Cancer</td>
<td>120436</td>
<td>MSH2</td>
<td>U03911 9.2e-261</td>
<td>M84170</td>
<td>DNA repair protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hereditary Non-polyposis Colon Cancer</td>
<td>120436</td>
<td>MLH1</td>
<td>U07418 6.3e-196</td>
<td>U07187</td>
<td>DNA repair protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cystic Fibrosis</td>
<td>219700</td>
<td>CFT2</td>
<td>M28668 1.3e-167</td>
<td>L35237</td>
<td>Metal resistance protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilson Disease</td>
<td>277900</td>
<td>WND</td>
<td>U11700 5.9e-161</td>
<td>L36317</td>
<td>Probable copper transporter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycerol Kinase Deficiency</td>
<td>307030</td>
<td>GK</td>
<td>L13943 1.8e-129</td>
<td>X69049</td>
<td>Glycerol kinase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bloom Syndrome</td>
<td>210900</td>
<td>BLM</td>
<td>U02817 2.6e-119</td>
<td>U22341</td>
<td>Helicase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adrenoleukodystrophy, X-linked</td>
<td>300100</td>
<td>ALD</td>
<td>U21876 9.4e-107</td>
<td>U17065</td>
<td>Peroxisomal ABC transporter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ataxia Telangiectasia</td>
<td>208900</td>
<td>ATM</td>
<td>U24655 2.8e-90</td>
<td>U31331</td>
<td>PI3 kinase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amyotrophic Lateral Sclerosis</td>
<td>105400</td>
<td>SOD1</td>
<td>K00065 2.0e-58</td>
<td>J03279</td>
<td>Superoxide dismutase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myotonic Dystrophy</td>
<td>160900</td>
<td>DM</td>
<td>L19268 5.4e-53</td>
<td>M21307</td>
<td>Serine/threonine protein kinase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lowe Syndrome</td>
<td>309000</td>
<td>OCRL</td>
<td>M88162 1.2e-47</td>
<td>Z47047</td>
<td>Putative IPP-5-phosphatase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurofibromatosis, Type 1</td>
<td>162200</td>
<td>NF1</td>
<td>M99914 2.0e-46</td>
<td>M33779</td>
<td>Inhibitory regulator protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choroideremia</td>
<td>303100</td>
<td>CHM</td>
<td>X78121 2.1e-42</td>
<td>S69371</td>
<td>GDP dissociation inhibitor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diastrophic Dysplasia</td>
<td>222600</td>
<td>DTD</td>
<td>U14528 7.2e-38</td>
<td>X82013</td>
<td>Sulfate permease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lissencephaly</td>
<td>247200</td>
<td>LIS1</td>
<td>L13385 1.7e-34</td>
<td>L26505</td>
<td>Methionine metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thomsen Disease</td>
<td>160800</td>
<td>CLC1</td>
<td>Z25884 7.9e-31</td>
<td>Z23117</td>
<td>Voltage-gated chloride channel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilms Tumor</td>
<td>194070</td>
<td>WT1</td>
<td>X51630 1.1e-20</td>
<td>X67787</td>
<td>Sulphite resistance protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achondroplasia</td>
<td>100800</td>
<td>FGFR3</td>
<td>M58051 2.0e-18</td>
<td>U7163</td>
<td>Serine/threonine protein kinase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menkes Syndrome</td>
<td>309400</td>
<td>MNK</td>
<td>X69208 2.1e-17</td>
<td>L36317</td>
<td>Probable copper transporter</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Major Application II: Finding Homologues (cont.)

• Cross-Referencing, one thing to another thing
• Sequence Comparison and Scoring
• Analogous Problems for Structure Comparison
• Comparison has two parts:
  (1) Optimally Aligning 2 entities to get a Comparison Score
  (2) Assessing Significance of this score in a given Context

• Integrated Presentation
  ◦ Align Sequences
  ◦ Align Structures
  ◦ Score in a Uniform Framework
Major Application II: Overall Genome Characterization

• Overall Occurrence of a Certain Feature in the Genome
  ◊ e.g. how many kinases in Yeast

• Compare Organisms and Tissues
  ◊ Expression levels in Cancerous vs Normal Tissues

• Databases, Statistics

(Clock figures, yeast v. Synechocystis, adapted from GeneQuiz Web Page, Sander Group, EBI)
Simplifying Genomes with Folds, Pathways, &c

~100000 genes

~1000 folds

~1000 genes

(human)

(T. pallidum)
At What Structural Resolution Are Organisms Different?

<table>
<thead>
<tr>
<th>person plant</th>
<th>protein fold (Ig)</th>
<th>super-secondary structure (ββ, TM–TM, αβαβ, ααα)</th>
<th>helix</th>
<th>individual atom (C,H,O...)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="person" /></td>
<td><img src="image" alt="protein fold" /></td>
<td><img src="image" alt="super-secondary structure" /></td>
<td><img src="image" alt="helix" /></td>
<td><img src="image" alt="individual atom" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1m</th>
<th>100Å</th>
<th>10Å</th>
<th>1Å</th>
</tr>
</thead>
</table>

### Practical Relevance

(Pathogen only folds as possible targets)

![Drug](image)